Regression analysis

MATH 5358

Department of mathematics, UTA

Course information

- Course Webpage: Please find a link in Blackboard.
- Office Hours: PKH 446, Mo/We 4:30-5:30 pm or by appointment
- Prerequisite: MATH 5312 or MATH 5305 with a B or better. Basic programming skills are preferred, but not required.
- Required Textbooks
  - Sheather, S. (2009). A Modern Approach to Regression with R. Springer.

- Other Recommended Textbooks and Resources
  - Chatterjee, S. and Hadi, A. S. (2012). Regression Analysis by Example. John Wiley & Sons.
  - Weisberg, S. (2013). Applied Linear Regression. Wiley.
  - James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R. Springer.
  - Gelman, A. and Hill, J. (2007). Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.

Coursework

- Homework assignments (35%)
  - 9 HW assignments (30 each) and a final HW (50).
  - You shall get 100% for homework if you earn no less than 90% of homework.
- Suppose your total is x points, then your final score for homework will be calculated as

$$100 \times \min\left(\frac{x}{0.9 \times 320}, 1\right)$$

- Midterm 1 (20%)
  - In-class exam.
  - A hand-written cheatsheet is allowed.
    - single page, letter size
- Midterm 2 (20%)
  - Take-home exam (48 hours)
- Final projects (25%)
  - Real-life applications or Statistical methodologies
  - The final project guideline will be announced after 1st midterm.

Assignment submission format

- A single homework assignment may be mixed with two types of assignments:
  - 1) Theoretical assignments will be tagged with (*Written*),
  - 2) R assignments will require the use of statistical software R.
- The R assignments must be turned in electronically, through Blackboard.
  - These should be be submitted in R Markdown and PDF format (two files).
  - Work submitted in R Markdown format that cannot be compiled, i.e., fails “Knit PDF”, will receive an automatic grade of 0.

Assignment submission format

- For the theoretical assignments, you can
  - 1) use R Markdown to include math equations in PDF,
  - 2) submit the assignments to the instructor in class, or
  - 3) scan them and turn in electronically, through Blackboard.
- The homework problems will be assigned weekly on the course webpage.
- It is your responsibility to check the course website regularly (at least once a day). Don't forget to refresh the site!

- It is encouraged to discuss homework problems with classmates or the instructor but you should finish your homework independently.

- Blackboard runs plagiarism detection software on all your submissions.

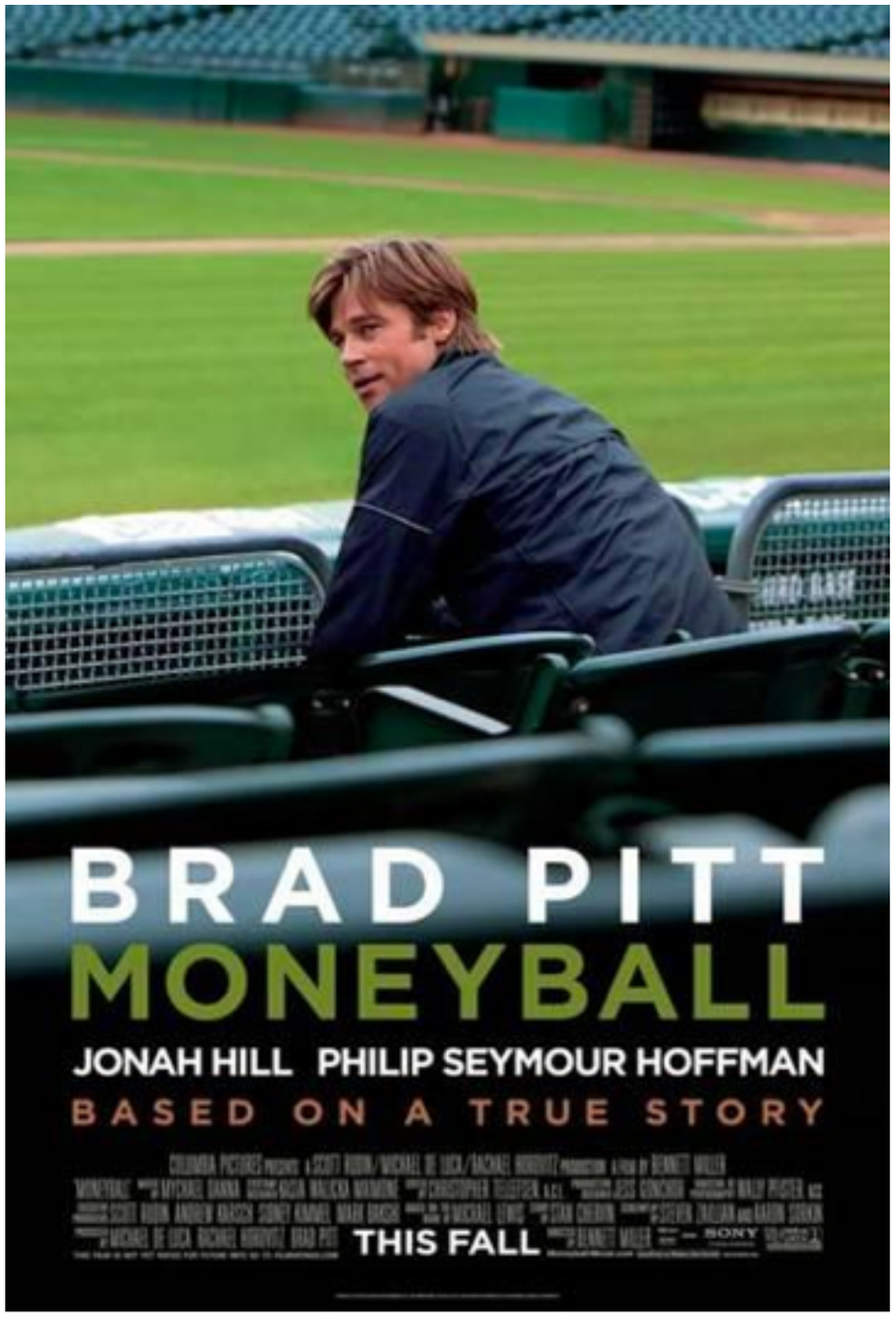
- For the take-home midterm, you cannot discuss problems with anybody except the instructor.

- Copying solutions from another student is cheating and plagiarism, and is a violation of Academic Integrity.

Students should

- be comfortable with the following concepts:
  - probability distribution functions, expectations, variance, conditional distributions, hypothesis testing, p-value, confidence interval, vector, matrix, matrix multiplication.
- review lecture slides and lab materials.
- finish reading assignments (see course schedule).
- write answers for coursework on your own codes and words.

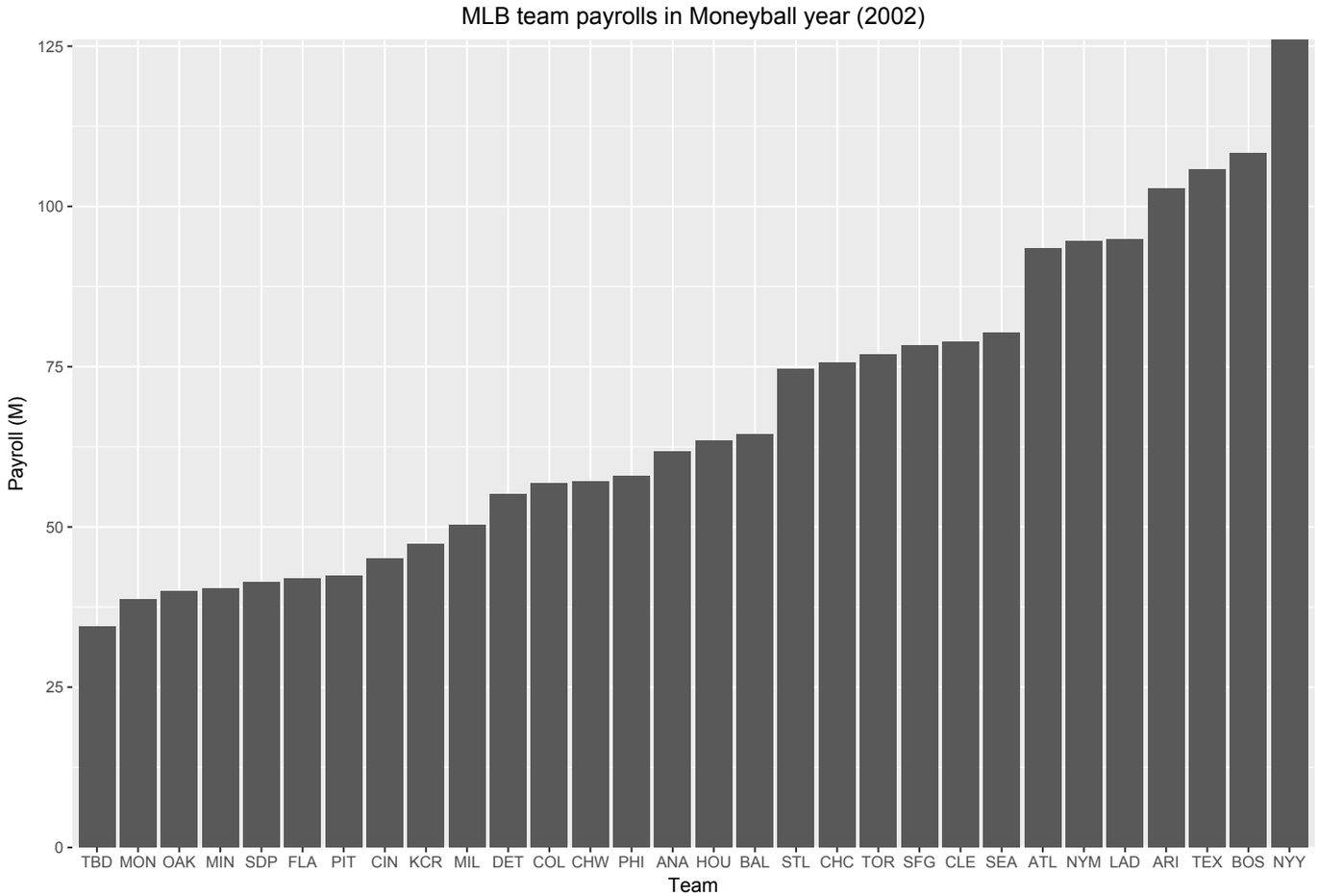
Moneyball



- Story about Oakland Athletics (A's) in 2002.
- How data-driven modeling can be used for better decision making and for winning.
- Collective wisdom of baseball insiders may be subjective and often flawed.

How to win MLB?

Baseball is an unfair game

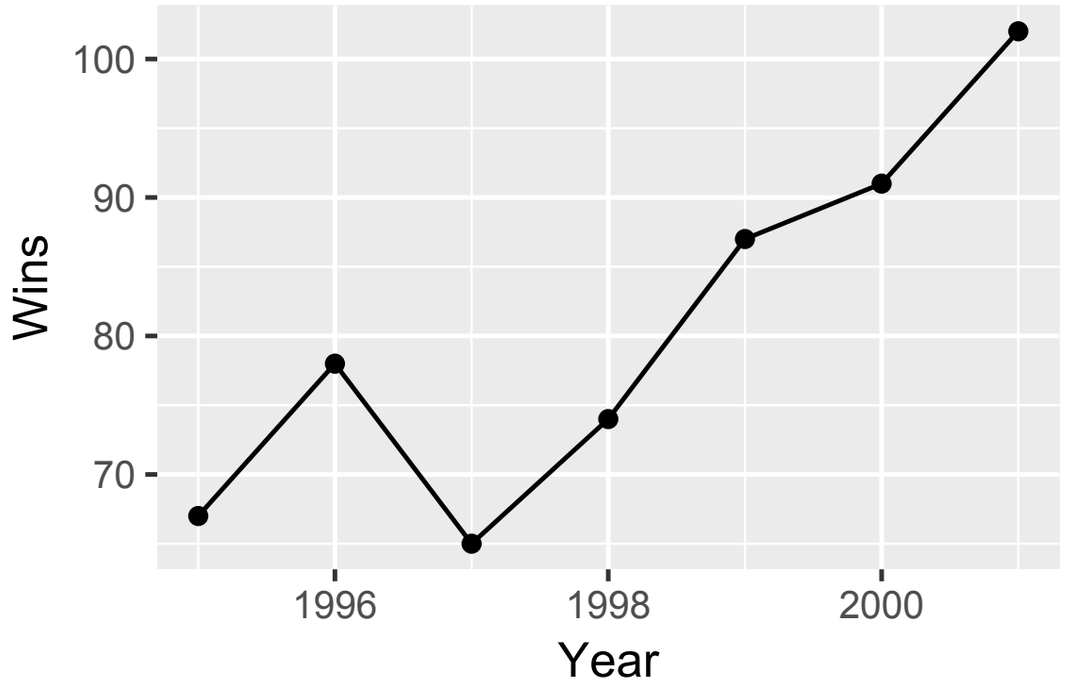


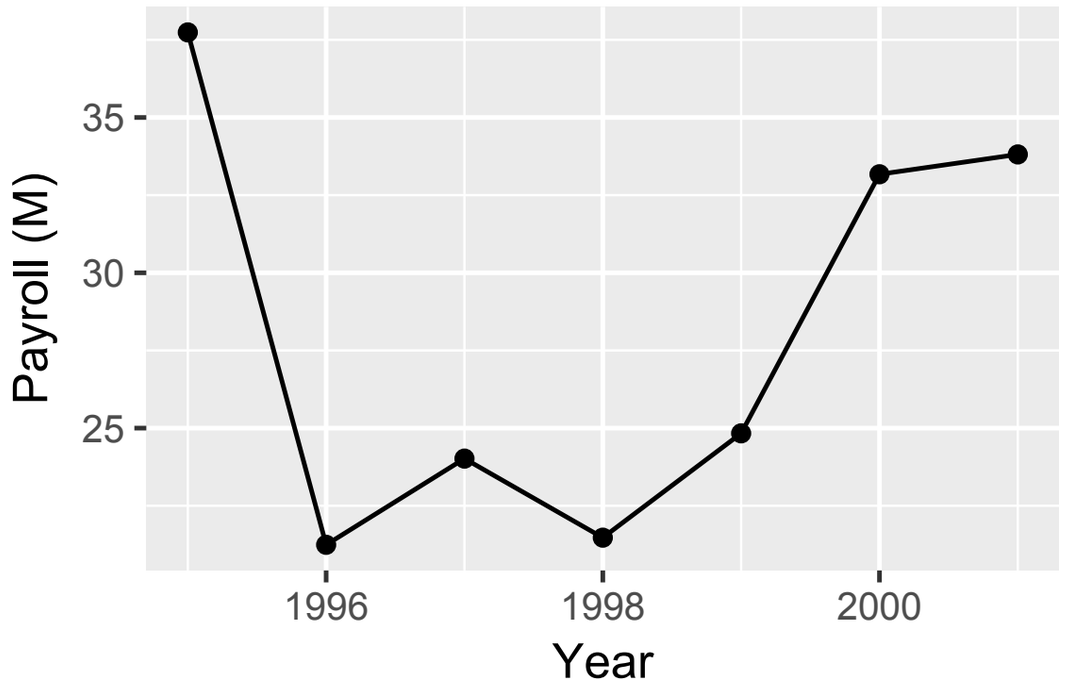
- Money is a very important aspect in every professional sport.
- A's are a small market team.
  - cannot meet salary demands of all star players.

Cost per win

	Teams	M$ / Wins
17	Texas Rangers	0.467
18	Los Angeles Dodgers	0.503
22	Oakland Athletics	0.563
25	New York Yankees	0.618

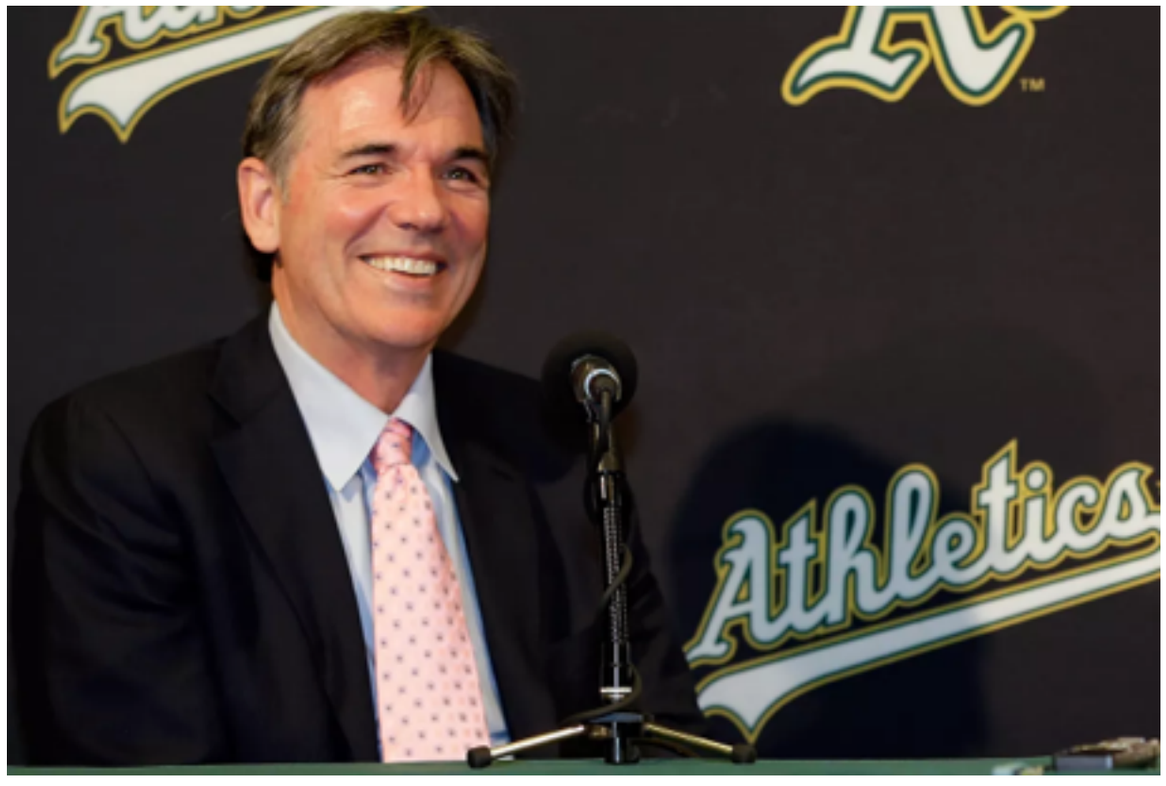
- In 1995, A's cost per win is as much as rich teams. A's failed to advance to playoffs.
  - A's owners ordered to slash payroll.
- In 1999 - 2001 seasons, A's earned 20+ wins with less spending than 1995





Billy Beane

- He was named general manager after the 1997 season
- Responsible for building a team
  - Controls player transactions
  - Hires/fires staffs
- Applied statistical analysis (sabermetrics) to players
- Led teams to reconsider how they evaluate players.



Paul DePodesta

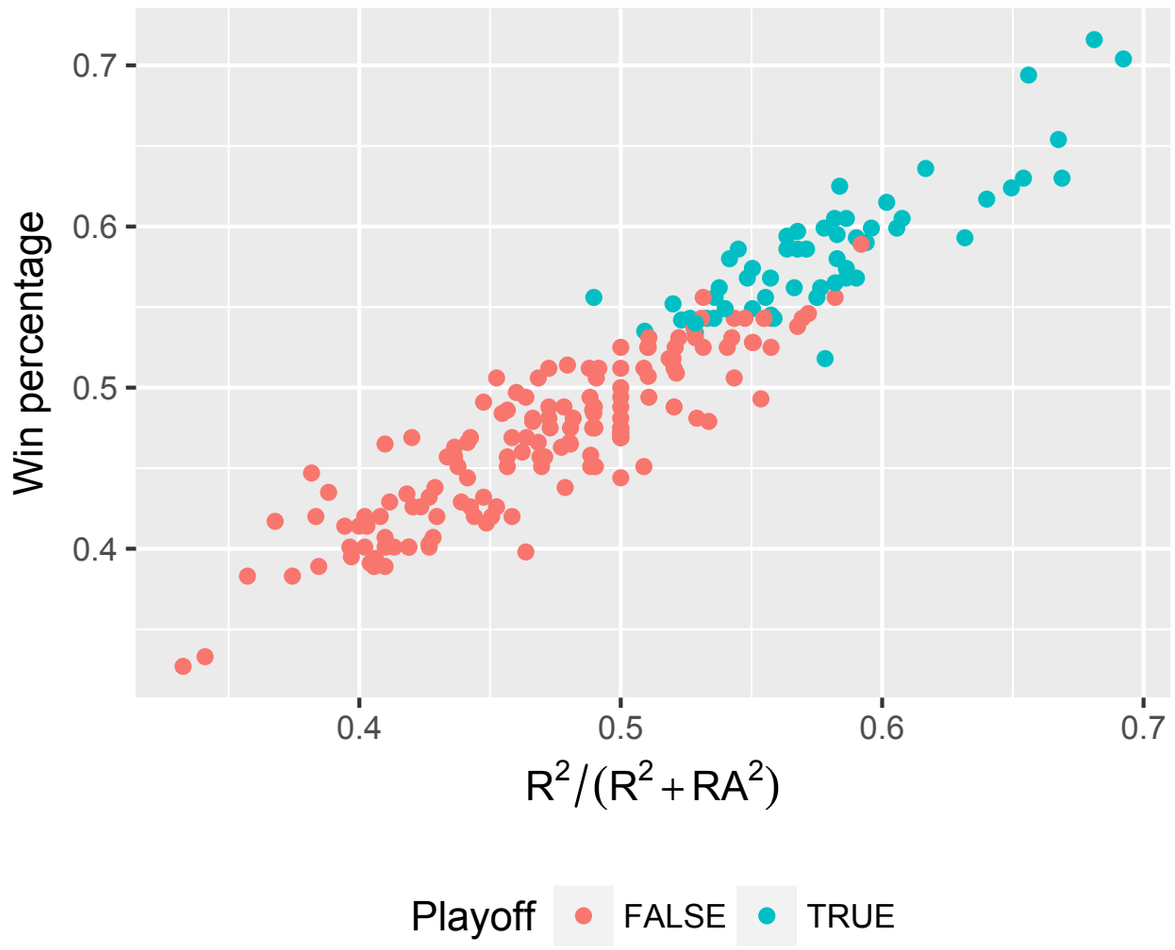
- Hired as a Billy Beane's assistant in 1999 (major in economics).
- A's lost three key players in 2001.
- A's needed to re-build a competitive roster on a limited budget.
- Paul DePodesta used a statistical model to list undervalued players who could become baseball stars.

Moneyball approaches

- Win Percentage 0.60+ to secure a playoff seat.
- Over the course of a season the A's should score at least 761 runs and not allow anymore than 618 runs.
- How to predict runs (R) and run allowed (RA)?

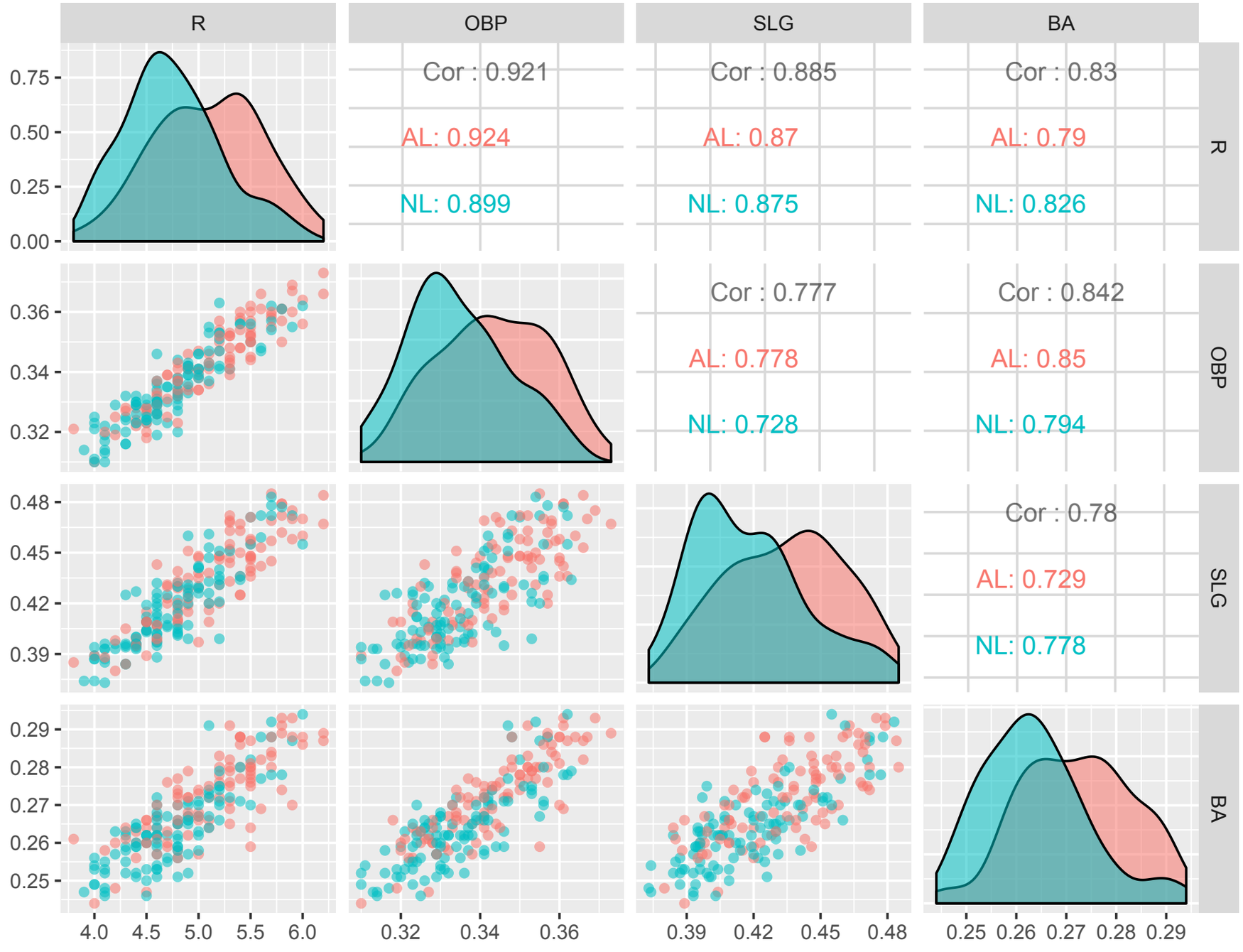
Pythagorean Theorem of Baseball

$$Win\% \approx \frac{R^2}{R^2 + RA^2}$$



Moneyball approach to predict R (runs)

- OBP = (H + BB + HBP)/(AB + BB + HBP + SF)
  - On-base percentage: how frequently a batter reaches base per plate appearance?
  - invented in the 1940s
- SLG = (1B + 2*2B + 3*3B + 4*HR) / AB)]
  - Slugging percentage: a measure of the batting productivity of a batter.



$$R_{it} = \beta_0 + \beta_1 OBP_{it} + \beta_2 SLG_{it} + \varepsilon_{it}$$

Index: i - team, t - year

The last term is “luck”.

How to model runs

- A's found OBP is the most important metric followed by SLG.
- A's claimed BA (batting average) was an overrated, and (OBP, SLG) was underrated metrics.
- For 2002, A's built a prospect roster that can score 769 runs.
- Many of the A's scouts had serious doubts about their team.

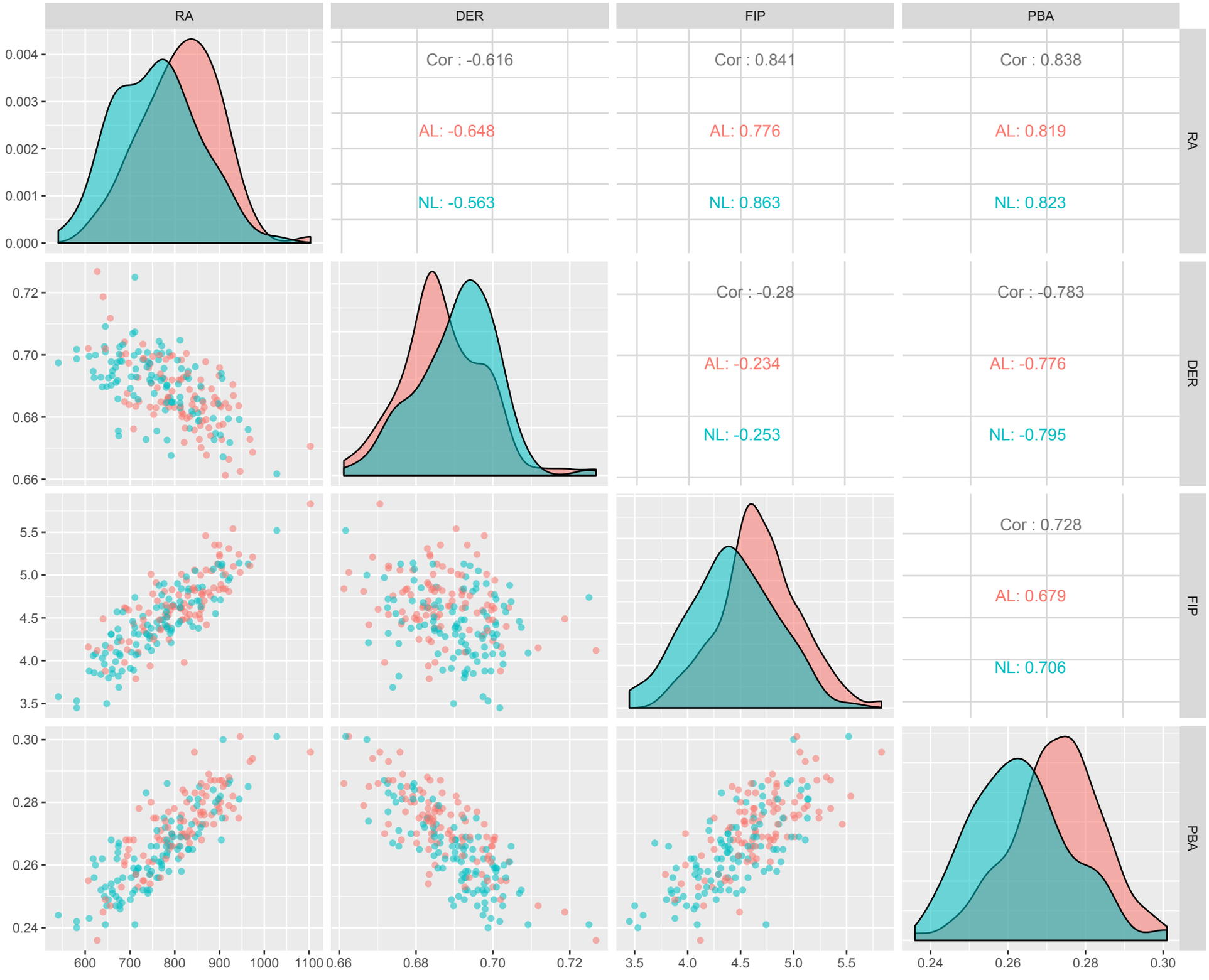
$$R_{it} = \beta_0 + \beta_1 OBP_{it} + \beta_2 SLG_{it} + \varepsilon_{it}$$

	Estimate	Std. Error	t
(Intercept)	-952.59	64.62	-14.74
OBP	3098.74	296.53	10.45
SLG	1616.01	152.53	10.60

R-squared: 0.8316

Moneyball approach to predict RA (runs allowed)

- Good defense = good fielding + good pitching
- DER = 1 - ((H + ROE - HR)/(PA-BB-SO-HBP-HR))
  - Defensive efficiency ratio (team defense): Higher is better.
    - For every ball hit into the field of play, how likely is the defense to convert that into an out?
- FIP = ((13*HR) + (3*(BB+HBP)) - (2*SO)) / IP + cFIP
  - Fielding independent pitching (pitcher performance): Lower is better.
  - FIP measures the events that are directly under a pitcher's control: strikeouts, walks, and home runs.



$$RA_{it} = \beta_0 + \beta_1 DER_{it} + \beta_2 FIP_{it} + \varepsilon_{it}$$

How to model runs allowed

- A's found FIP is the most important metric followed by DER.
- A's claimed PBA (batting average against pitcher) was an overrated, and (DER, FIP) was underrated metrics.
- For 2002, A's built a prospect roster that may allow 618 runs.
- Again, many of the A's scouts had serious doubts about their team.

$$RA_{it} = \beta_0 + \beta_1 DER_{it} + \beta_2 FIP_{it} + \varepsilon_{it}$$

	Estimate	Std. Error	t
(Intercept)	2497.870	171.636	14.55
DER	-3572.604	234.709	-15.22
FIP	164.724	6.146	26.80

R-squared: 0.8642

Prediction for 2002

OBP	SLG	DER	FIP
0.339	0.432	0.705	3.87

Pythagorean Theorem of Baseball

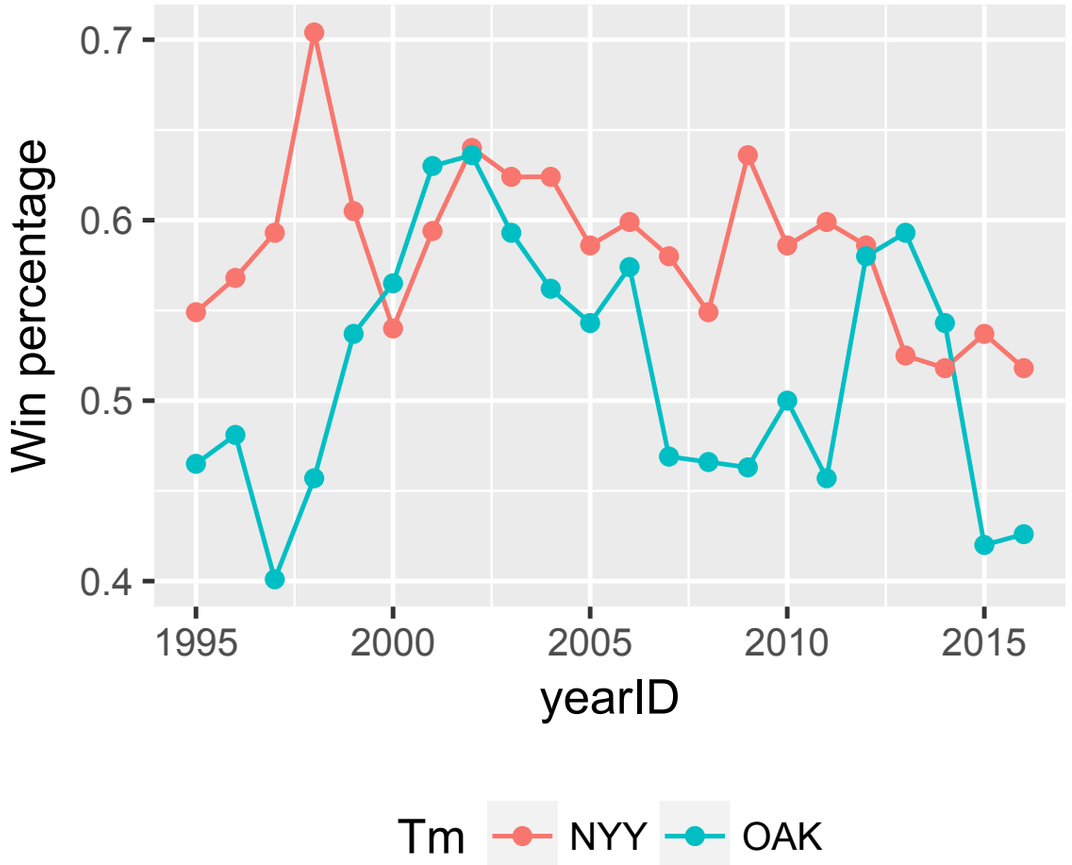
$$Win\% \approx \frac{769^2}{796^2 + 618^2} = 0.624$$

- Predicted Win% for 2002 was 62.4%.
- A's actual Win% in 2002 was 63.6%

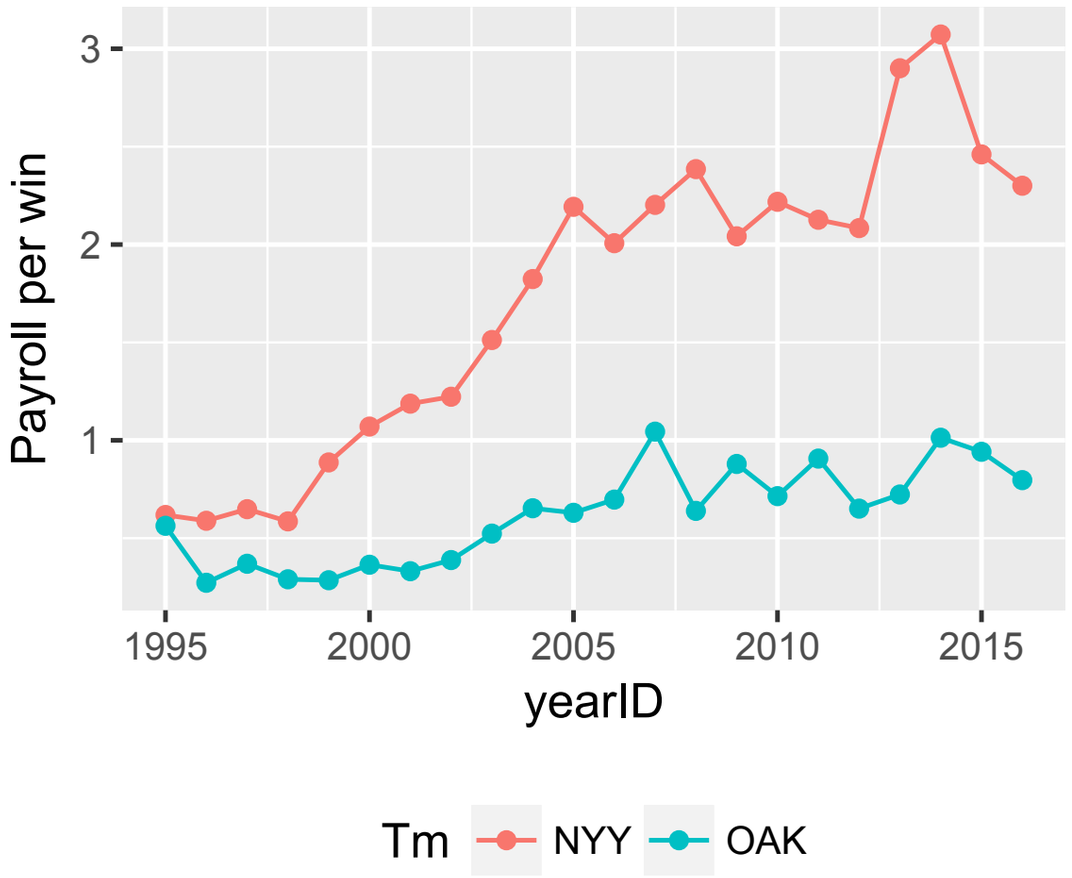
	Estimate	Std. Error	t
(Intercept)	-952.59	64.62	-14.74
OBP	3098.74	296.53	10.45
SLG	1616.01	152.53	10.60

	Estimate	Std. Error	t
(Intercept)	2497.870	171.636	14.55
DER	-3572.604	234.709	-15.22
FIP	164.724	6.146	26.80

- In 2002, A's won 20 consecutive games.
- A's advanced to Playoffs in 2002-3, 2006, 12-14), but A's never made it.
- A's just got unlucky in the playoffs



- Luck in MLB even out over the regular season (162 games played).
- In postseasons, the number of games is not large enough to even out good or bad luck.



Remarks

- A's achieved a great success using the moneyball techniques.
- A's ticket sales are not as good as a large market team such as NYY. People like to see star players, and their market is not as large as NYY.
- Boston Red Sox used Sabermetrics, and won the world series in 2007.
- Some models have been simplified. You can find Sabermetrics in http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.457.3155&rep=rep1&type=pdf

Over the semester, you will

- Visually explore data and manipulate data sets.
- Calculate/interpret linear models and inferential statistics for
  - Continuous responses (linear regression)
  - Discrete responses (logistic regression)
  - Correlated responses (AR(1) model)
- Perform various model diagnostics and improve the model.
- Infer appropriate conclusions about populations based on data.
- Compare and contrast various models / select a model.
- Be proficient with statistical software such as the R language.
- Understand how to create reproducible research documents using the R Markdown syntax.